PEMODELAN PRINCIPAL COMPONENT REGRESSION DENGAN SOFTWARE R

Margaretha Ohyver

Mathematics & Statistics Department, School of Computer Science, Binus University
Jln. K.H. Syahdan No. 9, Palmerah, Jakarta Barat 11480
mohyver@binus.edu; e_ohyver@yahoo.co.id

ABSTRACT

Principal Component Regression (PCR) is one method to handle multicollinear problems. PCR produces principal components that have a VIF less than ten. The purpose for this research is to obtained PCR model using R software. The result is a model of PCR with two principal components and determination coefficients $R^2 = 97{,}27\%$.

Keywords*: multicollinear, principal component regression, R software.*

ABSTRAK

Principal Component Regression (PCR) merupakan salah satu metode yang dapat digunakan untuk mengatasi masalah multikolinear. PCR menghasilkan komponen-komponen utama yang memiliki VIF kurang dari sepuluh. Tujuan dari penelitian ini adalah untuk memperoleh model PCR dari data yang mengandung multikolinear dengan bantuan software ***R****. Hasil yang diperoleh adalah model PCR dengan dua komponen utama dan koefisien determinasi* $R^2 = 97{,}27\%$.

Kata kunci*: multikolinear, principal component regression, software R.*

PENDAHULUAN

Salah satu metode statistika yang sering digunakan untuk menyelesaikan permasalahan adalah regresi. Metode ini digunakan untuk menganalisis hubungan antar variabel yang dinyatakan dalam sebuah persamaan yang disebut persamaan regresi. Ada dua variabel yang terlibat dalam persamaan ini, yaitu variabel bebas (X) dan variabel respon (Y). Apabila persamaan regresi memuat satu variabel bebas (X), model regresinya disebut model regresi sederhana. Apabila persamaan regresi memuat lebih dari satu variabel bebas (X), model regresinya disebut model regresi ganda.

Seperti halnya metode statistika lainnya, model regresi ganda mempunyai beberapa asumsi. Salah satu asumsinya adalah tidak terjadi multikolinear. Yang dimaksud dengan multikolinear adalah adanya korelasi antar variabel bebas. Adanya kasus ini dapat menyebabkan sulitnya memisahkan pengaruh masing-masing variabel bebas (X) terhadap variabel respon (Y). Asumsi yang terakhir sering terjadi pada data yang diambil dari keadaan tak terkontrol. Multikolinear juga dapat menyebabkan kesalahan tanda (positif atau negatif) dari dugaan koefisien regresi kuadrat terkecil. Akibat adanya pengaruh yang ditimbulkan oleh multikolinear tersebut, maka diperlukan suatu metode untuk mengatasinya

Ada beberapa metode yang dapat digunakan untuk mengatasi multikolinear, di antaranya *Partial Least Squares* (PLS), regresi ridge, dan *Principal Component Regression* (PCR). Aplikasi PLS dapat dilihat pada Ohyver (2010: 39-47). PLS digunakan pada data gingerol. Berdasarkan penelitian diketahui bahwa dengan menggunakan dua komponen diperoleh $R^2 = 83{,}8\%$ dan $Root\ Mean\ Squared\ Error$ (RMSE) $= 0{,}100891$. Pemodelan PLS dilakukan dengan menggunakan Minitab. Aplikasi regresi ridge dapat dilihat pada Ohyver (2011: 451-457). Regresi ridge digunakan untuk memodelkan hubungan antara 6 (enam) variabel bebas yang digunakan, yaitu: X_1 adalah benih (*ml*), X_2 adalah pupuk urea (*kg*), X_3 adalah pupuk TSP (*kg*), X_4 adalah pupuk KCL (*ml*), X_5 adalah pestisida (*ml*), X_6 adalah curahan hari kerja (*HKP*), Y adalah hasil produksi (*kg*). Ohyver menggunakan *software* NCSS. Aplikasi PCR dapat dilihat pada Silalahi (2011). Pada pemodelan PCR, Silalahi menggunakan *software* SPSS. Pada makalah ini akan dibahas pemodelan dengan PCR pada data sekunder yang diperoleh dari Pradipta (2009).

PCR merupakan salah satu metode yang dapat digunakan untuk mengatasi masalah multikolinier. Metode ini akan menghasilkan komponen-komponen utama yang tidak berkorelasi. Yang perlu dicatat adalah jika semua komponen utama diikut sertakan dalam model regresi, maka model yang dihasilkan akan sama dengan yang diperoleh dengan metode kuadrat terkecil. Jika hanya beberapa komponen utama saja yang diikut sertakan, maka akan diperoleh penduga koefisien regresi yang bias namun memiliki *variance* yang minimum (Jollife, 2002).

Pemodelan PCR pada makalah ini akan dilakukan dengan menggunakan bantuan *software*. Secara umum ada dua macam kelompok paket *software* statistik. Dua kelompok tersebut adalah kelompok *software* komersil dan kelompok *software* statistik *open source* atau *freeware*. *Software* yang termasuk dalam kelompok pertama antara lain MINITAB, SPSS, dan SAS. Sedangkan *software* yang termasuk dalam kelompok kedua antara lain R dan Vista (Suhartono, 2008).

Software sangat memegang peranan penting untuk keperluan analisis data. Untuk menggunakan *software* kelompok pertama sangat dibutuhkan biaya yang relatif mahal bagi sebagian besar pengguna di Indonesia. Alternatif lain adalah menggunakan *software* kelompok kedua, yang salah satunya adalah R.

R adalah bahasa komputer yang memungkinkan pengguna dalam hal algoritma program dan menggunakan apa yang sudah dibuat oleh pengguna lain (Ohyver, 2011: 1). Pengguna dapat menuliskan fungsi-fungsi, melakukan kalkulasi, mengaplikasikan teknik-teknik statistika, menciptakan grafik sederhana dan rumit, dan bahkan membuat fungsi library milik sendiri. Kelebihan R dibanding beberapa *software* yang biasa digunakan oleh pengguna di Indonesia adalah *free of charge*. Untuk mengunduh dan menginstal R, pengguna dapat mengunjungi website www.r-project.org.

Seperti halnya *software* statistik yang lain, R juga dapat menjadi alat dalam analisis data. Mulai dari statistik deskriptif, analisis peluang, statistik multivariat, sampai analisis deret waktu. Pada makalah kali ini akan dilakukan pemodelan PCR dengan bantuan R. Sehingga permasalahan yang akan dibahas adalah bagaimana aplikasi PCR pada data yang mengandung multikolinear serta bagaimana aplikasi R dalam membantu pemodelan PCR. Adapun tujuan yang hendak dicapai adalah memperoleh model PCR untuk data yang mengandung multikolinear dengan bantuan R.

METODE

Data yang akan digunakan adalah data sekunder yang diperoleh dari Pradipta (2009), yang selanjutnya akan disebut data Pradipta. Ada tiga variabel bebas dan satu variabel respon yang terlibat. Variabel-variabel tersebut adalah barang impor (milliard Franc Perancis, y), barang yang dipesan (Milliard Franc Perancis X_1), persediaan barang (Milliard Franc Perancis), dan barang yang dikonsumsi (Milliard Franc Perancis, X_3). Dalam Pradipta (2009), data tersebut diolah dengan menggunakan regresi ridge. Dengan regresi ridge diperoleh koefisien determinasi (R^2) sebesar 93,42%.

Alasan penggunaan regresi ridge, PCR, dan berbagai metode yang lain, adalah adanya multikolinear. Yang dimaksud dengan multikolinear adalah adanya korelasi di antara variabel-variabel bebas dan hanya berlaku untuk hubungan linear. Adanya multikolinear dalam model regresi ganda dapat mengakibatkan *variance* dari $\hat{\boldsymbol{\beta}}$ membesar sehingga pengaruh masing-masing variabel bebas tidak dapat dipisahkan. Sehingga penambahan atau pengeluaran suatu variabel bebas akan mengubah koefisien regresi.

Multikolinear dapat dideteksi dengan analisis korelasi. Akan tetapi cara pendeteksian seperti ini tidak efektif apabila multikolinear terjadi di antara lebih dari dua variabel bebas. Sebagai contoh, antara X_1 dan X_2 berkorelasi rendah, tetapi antara X_1 dan X_2 terhadap X_3 berkorelasi tinggi. Suatu metode formal untuk mendeteksi adanya multikolinear adalah dengan *Variance Inflation Factor* (VIF). VIF mengukur seberapa besar variance koefisien regresi dugaan membesar dibandingkan variabel-variabel bebasnya tidak berkorelasi linear. Nilai VIF diperoleh dari persamaan berikut.

$$VIF_j = \frac{1}{1-R_j^2}, \quad j = 1, 2, \dots, p \tag{1}$$

R_j^2 adalah koefisien determinasi jika X_j diregresikan terhadap $(p-1)$ variabel lainnya di dalam model. Nilai VIF yang lebih besar dari sepuluh dapat dijadikan indikasi bahwa multikolinear telah mempengaruhi nilai dugaan kuadrat terkecil.

Model regresi secara umum dapat dituliskan seperti pada persamaan (2). Dimana $\mathbf{y}$ adalah vektor dari variabel respon untuk *n* pengamatan, $\mathbf{X}$ adalah matriks berukuran $(n \times p)$ yang elemen (i, j) adalah nilai dari variabel bebas ke-j untuk pengamatan ke-i, $\boldsymbol{\beta}$ adalah vektor dari p koefisien regresi dan $\boldsymbol{\epsilon}$ adalah vektor dari error.

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \tag{2}$$

Nilai-nilai dari komponen utama untuk setiap pengamatan dapat diperoleh dengan mengunakan persamaan (3). Dimana $\mathbf{Z}$ adalah nilai (skor) dari komponen utama (PC) ke-*k* untuk pengamatan ke-*i*, dan $\mathbf{A}$ adalah matriks berukuran $(p \times p)$ dengan kolom ke-*k* adalah vektor eigen ke-*k* dari $\mathbf{X^T X}$..

$$\mathbf{Z} = \mathbf{XA} \tag{3}$$

Karena $\mathbf{A}$ matriks ortogonal, $\mathbf{X\beta}$ dapat dituliskan menjadi $\mathbf{XAA^T\beta} = \mathbf{Z\gamma}$, dimana $\boldsymbol{\gamma}=\mathbf{A^T\beta}$. Persamaan (2) dapat dituliskan menjadi persamaan (4).

$$\mathbf{y} = \mathbf{Z\gamma} + \boldsymbol{\epsilon} \tag{4}$$

Atau

$$\mathbf{y} = \boldsymbol{Z}_m\boldsymbol{\gamma}_m + \boldsymbol{\epsilon}_m \tag{5}$$

Dimana $\boldsymbol{\gamma}_m$ adalah vektor dari m elemen yang merupakan subset dari elemen-elemen $\boldsymbol{\gamma}$, $\mathbf{Z}_m$ adalah matriks berukuran $(n \times m)$ yang kolomnya adalah subset korespondensi dari kolom-kolom $\mathbf{Z}$, dan ϵ_m adalah vektor error. Dengan menggunakan metode kuadrat terkecil, akan diperoleh koefisien regresi sebagai berikut.

$$\hat{\boldsymbol{\beta}} = \mathbf{A}\hat{\boldsymbol{\gamma}} \tag{6}$$

Pada PCR, variabel bebas (X) yang digunakan adalah variabel bebas yang dibakukan dan diskalakan. Variabel bebas tersebut diperoleh dengan menggunakan persamaan (4).

$$X^* = \frac{(X_{ji} - \bar{X}_j)}{\sqrt{\sum_{i=1}^{n}(X_{ji} - \bar{X}_j)^2}} \tag{7}$$

PCR menghasilkan komponen-komponen utama yang sudah tidak berkorelasi. Pemilihan jumlah komponen dapat dilakukan dengan memperhatikan PRESS, Cp, atau *variance*. Pada makalah ini, pemilihan jumlah komponen yang akan digunakan dilakukan dengan memperhatikan kontribusi komponen tersebut terhadap variabel respon (Y). Penelitian ini akan dilakukan dengan langkah-langkah sebagai berikut. Pertama, membuat persamaan regresi ganda. Kedua, menghitung VIF dengan menggunakan persamaan. Ketiga, membuat persamaan PCR.

HASIL DAN PEMBAHASAN

Data yang digunakan adalah data yang sebelumnya telah dimodelkan dengan menggunakan regresi ridge. Sehingga sudah pasti ada multikolinear. Akan tetapi karena makalah ini juga membahas tentang R maka tetap akan ditunjukkan adanya multikolinear serta pemodelan dengan regresi ganda. Pembentukan model regresi ganda akan dibantu dengan *software* **R**, dalam hal ini **R Commander**. Analisis regresi dapat dilakukan melalui menu **Statistics**, kemudian pilih **Fit models**, dan pilih **Linear regression**. Kemudian akan muncul jendela dialog yang dapat dilihat pada Gambar 1.

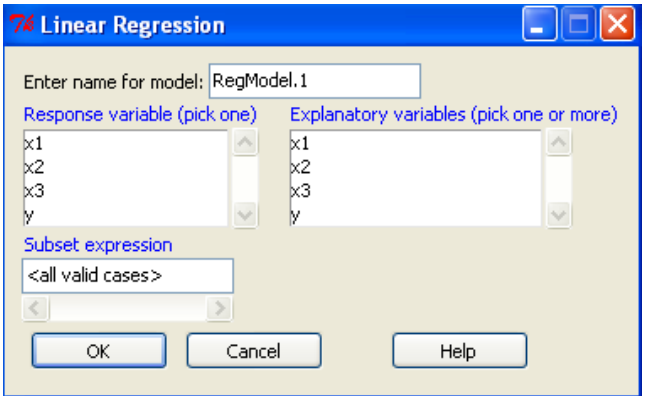
Setelah memasukkan seluruh variabel bebas dan variabel responnya, diperoleh *output* seperti yang terdapat pada Gambar 2. Persamaan regresinya dapat dilihat pada persamaan (8).

$$y = -19{,}86 + 0.03X_1 + 0{,}43X_2 + 0{,}24X_3 \tag{8}$$

Berdasarkan nilai *p-value* diketahui bahwa variabel X_1, X_2, dan X_3 secara parsial tidak berpengaruh signifikan. Padahal ketiga variabel secara logika harusnya mempengaruhi nilai variabel y. Selain itu berdasarkan *p-value* untuk uji F, diperoleh hasil bahwa paling sedikit ada satu variabel yang berpengaruh secara signifikan.

Jika hanya dua variabel bebas, yaitu X_1 dan X_2 yang digunakan, *output*-nya dapat dilihat pada Gambar 3. Persamaan regresinya dapat dilihat pada persamaan (9).

$$y = -16{,}89 + 0{,}19X_1 + 0{,}42X_2 \tag{9}$$



Gambar 1. Jendela dialog regresi linear.

Berdasarkan nilai *p-value* diketahui bahwa hanya variabel X_2 yang berpengaruh secara signifikan. Sedangkan *p-value* untuk uji F diperoleh hasil bahwa paling sedikit ada satu variabel yang berpengaruh secara signifikan.

Jika dibandingkan antara Gambar 2 dan Gambar 3, diketahui bahwa nilai-nilai koefisien dugaannya berbeda. Misal koefisien dugaan untuk X_1 di Gambar 2 adalah 0, 03133 sedangkan koefisien dugaan untuk X_1 di Gambar 3 adalah 0, 191278. Demikian pula untuk X_2.

```
Residuals:
    Min      1Q  Median      3Q     Max
-2.7693 -1.8761 -0.3555  1.3489  4.0391

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -19.85806    4.14574  -4.790 0.000288 ***
x1            0.03133    0.18781   0.167 0.869900
x2            0.43144    0.32386   1.332 0.204077
x3            0.24450    0.28678   0.853 0.408239
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.269 on 14 degrees of freedom
Multiple R-squared: 0.9729, Adjusted R-squared: 0.9671
F-statistic: 167.5 on 3 and 14 DF,  p-value: 3.34e-11
```

Gambar 2. Output tiga variabel bebas.

Berdasarkan hal tersebut di atas dapat dicurigai adanya kasus multikolinear. Untuk itu perlu dilakukan pengecekan multikolinear dengan mengecek nilai VIF. Dengan menu **Models > Numerical diagnostics > Variance-inlation factors**, akan diperoleh *output* pada Gambar 4. Pada Gambar 4 diketahui bahwa ada nilai yang lebih dari 10, yaitu 469, 742135. Berdasarkan hal ini maka selanjutnya data akan dimodelkan dengan menggunakan regresi komponen utama.

Pembentukan model regresi komponen utama akan dibantu dengan *software* **R**, dalam hal ini **R *Commander***. Analisis komponen utama dapat dilakukan melalui menu ***Statistics***, kemudian pilih ***Dimensional analysis***, dan pilih ***Principal-components analysis***. Kemudian akan muncul jendela dialog yang dapat dilihat pada Gambar 5.

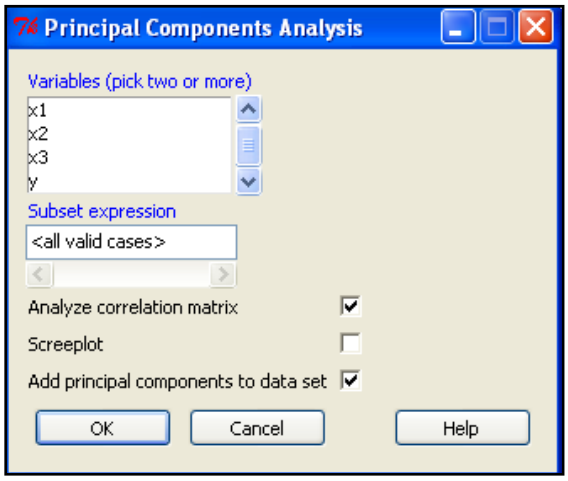
Setelah memasukkan variabel-variabelnya dan memilih ***Analyze correlation matrix*** dan ***Add principal components to data set***, akan diperoleh *output* yang ditampilkan pada Gambar 6.

```
Residuals:
    Min      1Q  Median      3Q     Max
-3.7238 -1.6075  0.2222  1.3097  3.8970

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -16.889560   2.229861  -7.574 1.68e-06 ***
x1            0.191278   0.008793  21.754 9.30e-13 ***
x2            0.422105   0.320715   1.316    0.208
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.249 on 15 degrees of freedom
Multiple R-squared: 0.9715, Adjusted R-squared: 0.9677
F-statistic: 255.5 on 2 and 15 DF,  p-value: 2.589e-12
```

Gambar 3. Output dua variabel bebas.

```
> vif(RegModel.2)
        x1         x2         x3
469.742135   1.049877 469.371343
```

Gambar 4. Nilai VIF.



Gambar 5. Jendela dialog *principal component analysis.*

Gambar 6 menunjukkan *output* sebagai berikut. *Component loadings* adalah vektor eigen yang persamaannya dapat dilihat pada persamaan 10.

$$\left(\mathbf{X}^{*\mathbf{T}}\mathbf{X}^{*} - \lambda_j \mathbf{I}\right)\mathbf{V_j} = \mathbf{0} \tag{10}$$

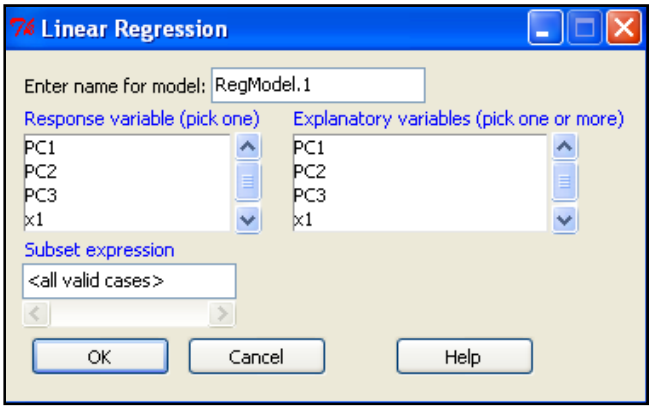
λ_j merupakan nilai eigen yang nilainya juga terdapat pada Gambar 6. *Component variances* adalah nilai-nilai eigen yang dimaksud. Untuk nilai-nilai komponen utama yang dihasilkan dapat dilihat pada Tabel 1. Jika Y diregresikan terhadap komponen-komponen utama yang ada pada Tabel 1 akan diperoleh hasil pada Gambar 3. Regresi dapat dilakukan dengan melalui menu ***Statistics***, kemudian pilih ***Fit models***, dan pilih ***Linear regression***. Kemudian akan muncul jendela seperti pada Gambar 7.

```
> unclass(loadings(.PC))  # component loadings
      Comp.1     Comp.2       Comp.3
x1 0.6810390  0.1897074  0.707246074
x2 0.2695946 -0.9629730 -0.001302688
x3 0.6808118  0.1915569 -0.706966261

> .PC$sd^2  # component variances
     Comp.1      Comp.2      Comp.3
2.083885517 0.915049056 0.001065427

> summary(.PC) # proportions of variance
Importance of components:
                          Comp.1    Comp.2       Comp.3
Standard deviation     1.4435669 0.9565820 0.0326408823
Proportion of Variance 0.6946285 0.3050164 0.0003551424
Cumulative Proportion  0.6946285 0.9996449 1.0000000000
```

Gambar 6. Output principal component analysis.

Tabel 1
Nilai-nilai Komponen Utama Data Pradipta

PC1	PC2	PC3
-1,889	-0,849	0,026
-1,661	-0,724	0,045
-1,565	-0,084	0,017
-1,458	-0,054	-0,002
-1,631	1,125	-0,031
-1,252	0,556	-0,016
-1,002	0,688	-0,031
-0,195	-1,234	-0,057
-0,001	-0,812	-0,043
0,113	-0,841	-0,012
-0,454	1,700	0,015
0,691	-0,986	0,068
0,715	0,064	0,033
1,014	0,638	0,018
1,670	-0,098	-0,047
2,427	-1,359	-0,006
1,789	2,021	0,009
2,687	0,248	0,012



Gambar 7. Jendela dialog regresi.

Ada tiga komponen utama yang terbentuk. Dari ketiganya, akan digunakan dua komponen utama dengan pertimbangan *cumulative proportion* > 70%. Jika dilihat pada Gambar 8, diketahui bahwa $R^2 = 97{,}27\%$. Selain itu, VIF yang diperoleh juga < 10. Berdasarkan *output* tersebut dapat dituliskan persamaan untuk komponen utamanya adalah sebagai berikut.

$$Y = 30{,}0944 + 8{,}2405PC1 + 1{,}5564PC2 \quad (11)$$

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  30.0944     0.5189  57.994  < 2e-16 ***
PC1           8.2405     0.3595  22.924 4.33e-13 ***
PC2           1.5564     0.5425   2.869   0.0117 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.202 on 15 degrees of freedom
Multiple R-squared: 0.9727, Adjusted R-squared: 0.969
F-statistic: 266.9 on 2 and 15 DF,  p-value: 1.886e-12
```

Gambar 8. Output regresi *Y* terhadap komponen-komponen utama.

Untuk mendapatkan persamaan regresi bagi data tersebut, perlu dilakukan transformasi ke variabel asal dengan menggunakan persamaan (12).

$$\boldsymbol{\beta} = \mathbf{V}\boldsymbol{\alpha} \quad (12)$$

Dimana $\boldsymbol{\beta}$ adalah vektor dari koefisien regresi untuk variabel bebas asal, $\mathbf{V}$ adalah vektor eigen, dan $\boldsymbol{\alpha}$ adalah vektor dari koefisien regresi untuk variabel bebas yang baru. Berdasarkan persamaan (12) diperoleh nilai $\boldsymbol{\beta}$ sebagai berikut.

$$\boldsymbol{\beta} = \begin{bmatrix} 5{,}907362 \\ 0{,}722823 \\ 5{,}908369 \end{bmatrix}$$

Sehingga persamaan regresi untuk data Pradipta adalah sebagai berikut.

$$Y = 30{,}09 + 5{,}91X_1 + 0{,}72X_2 + 5{,}91X_3 \quad (13)$$

Berdasarkan persamaan (11) diketahui bahwa barang impor mendapat pengaruh positif dari barang yang dipesan, persediaan barang, dan barang yang dikonsumsi. Jika dibandingkan dengan metode yang digunakan oleh Pradipta (2009), hasil dengan PCR lebih baik jika ditinjau dari R^2.

PENUTUP

Berdasarkan hasil dan pembahasan sebelumnya, diperoleh kesimpulan sebagai berikut. Pertama, multikolinear yang ada pada data Pradipta dapat diatasi dengan menggunakan *Principal Component Regression*. Kedua, pemodelan *Principal Component Regression* dilakukan dengan menggunakan 2 komponen utama. Ketiga, *Software* **R** dapat digunakan untuk membantu pemodelan dengan *Principal Component Regression*.

DAFTAR PUSTAKA

Jollife, I. T. (2002). *Principal Component Analysis* (2nd ed). New York: Springer-Verlag.

Ohyver, M. (2010). Penerapan partial least squares pada data gingerol. *ComTech, 1*(1): 39-47.

Pradipta, N. (2009). *Metode Regresi Ridge untuk Mengatasi Model Regresi Linier Berganda yang Mengandung Multikolinearitas*. Skripsi tidak diterbitkan. Universitas Sumatera Utara, Medan. Diakses dari http://repository.usu.ac.id/bitstream/123456789/14037/1/09E01589.pdf.

Suhartono. (2008). *Analisis Data Statistik Dengan R*. Yogyakarta: Graha Ilmu.